Age Recommendation for Texts

Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, Aline Étienne

Age Recommendation for Texts

Alexis Blandin[1], **Gwénolé Lecorvé**[1], **Delphine Battistelli**[2], **Aline Étienne**[1,2]
[1]Univ Rennes, CNRS, IRISA Lannion, France
[2]Université Paris Nanterre, CNRS, MoDyCo Nanterre, France
{alexis.blandin, gwenole.lecorve}@irisa.fr, {delphine.battistelli, aline.etienne}@parisnanterre.fr

Abstract

The understanding of a text by a reader or listener is conditioned by the adequacy of the text's characteristics with the person's capacities and knowledge. This adequacy is critical in the case of a child since her/his cognitive and linguistic skills are still under development. Hence, in this paper, we present and study an original natural language processing (NLP) task which consists in predicting the age from which a text can be understood by someone. To do so, this paper first exhibits features derived from the psycholinguistic domain, as well as some coming from related NLP tasks. Then, we propose a set of neural network models and compare them on a dataset of French texts dedicated to young or adult audiences. To circumvent the lack of data, we study the idea to predict ages at the sentence level. The experiments first show that the sentence-based age recommendations can be efficiently merged to predict text-based recommendations. Then, we also demonstrate that the age predictions returned by our best model are better than those provided by psycholinguists. Finally, the paper investigates the impact of the various features used in these results.

Keywords: natural language processing, age recommendation, children, psycholinguistics, neural networks

1. Introduction

The way in which an individual understands a text is complex. It depends both on the characteristics of the text and the abilities of the reader or listener. For instance, important abilities are the capacities to remember information, to position an event in a story line, to analyze the structure of a sentence, to understand a word or simply to read a text. For a given person, these various aspects are function of her/his degree of neuro-cognitive development, linguistic mastering (including the ability to read) and culture. During childhood, all these aspects are highly and constantly evolving. Hence, when presenting a text to a child, it is crucial to assess if this text is adequate enough to be understood.

This paper aims at automatically predicting age recommendations for texts with the objective to maximize its understanding by a child. Potential applications are numerous. First, this may be helpful when one wants to provide texts to children, mainly in schools where teachers could be assisted when preparing their class, and in search engines where returned pages could be conditioned by the age of the user. Secondly, age recommendation could also be a precious tool for authors (be they professional or not) at present by analysing written drafts, or, in a further perspective, by proposing reformulations. Finally, the automatic study of features could give useful feedback to psycholinguists investigating children's text comprehension abilities, e.g. by providing them with linguistic criteria that will help them choose appropriate texts for children, interpret experiments' results, or find new avenues to explore.

Age recommendation can be broadly considered as a specific type of text readability task, that is the prediction of how difficult to read a text is for a specific population (François, 2015), e.g. is a text readable by a non-native person, or is this form readable for customers? However, few machine learning approaches have been proposed in text readability and, to our knowledge, none for age recommendation. Furthermore, text readability is centered on reading activity, whereas the current study is focused on language understanding and thus also includes orally transmitted texts, especially texts told to very young children. As such, we consider this task to be an original task. Our work contributes to it as follows:

1. First, while predicting an age can obviously be seen as a regression problem, it is not clear whether it is the most relevant approach since the age at which children acquire a given skill is usually variable. To investigate this issue, we implement several possible formalizations.

2. Secondly, contrary to popular NLP tasks like named entity recognition or machine translation, one cannot rely on massive textual corpora. To get around this problem, we suggest, on the one hand, collecting children-dedicated texts based on recommendations made by authors or editors. On the other hand, we introduce and evaluate the assumption that all sentences of a given textual document share the same recommendation as the whole text. That is, age recommendation is performed at the sentence level. While this assumption is obviously wrong, we show that it is sufficient to train efficient models at the sentence level and that aggregating results improves the prediction accuracy.

3. Finally, this paper provides first conclusions about what an acceptable error is by comparing our results with human performance. Likewise, we show that word embeddings contribute the most to the good results of the models, while additional features bring almost no extra improvement.

In this paper, Section 2. first browses the related work in psycholinguistics and computational linguistics. Based on this literature review, Section 3. presents the features selected for age prediction. Then, we introduce our various

prediction approaches in Section 4.. Finally, Section 5. details the dataset on which experiments are carried out while their results and the lessons than can be drawn from them are presented in Section 6..

2. Background and State of the Art

This section introduces some key results from psycholinguistics on the development of children's text understanding. A specific focus is given on how the learning to read process can be modeled. These constitute useful background knowledge since they may suggest some features or strategies for machine learning. Finally, related work are browsed in the field of computational linguistics.

2.1. Some Insights in Developmental Stages

The main reason why children do not understand texts like adults is that their brain is still developing. As an illustration, the peak of brain activity is at four, with an activity equivalent to 150% of that of an adult (Gathercole, 1999), when language acquisition reaches a key stage. In particular, short-term memory develops strongly between the ages of two and eight. This memory mainly affects the comprehension of language or complex task accomplishments (Gathercole, 1999). Since children first access language through speech, phonological short-term memory plays a decisive part in language comprehension. It affects various processes like the storage of acoustic information, the analysis and memorization of phonological information, a word repetition mechanism leading to long-term memorization, the recovery of stored information, and finally the linking of words heard to their morphosyntactic interpretation. The phonological aspects or the length of a sentence thus seem to be interesting criteria to study, since these elements directly involve phonological short-term memory.

Among semantic markers, as shown by (Tartas, 2010) or (Hickmann, 2012), the acquisition of temporal notions is crucial since it enables children to locate themselves in (calendar) time, as well as to chronologically order events. This acquisition follows key stages. From 0 to 1.5 years, only events in recent past can be properly ordered by children but the perception that an object or someone remains is active. Hence, cyclic stories with repetitions are more indicated. Then, from 1.5 to 4/5 years, children understand that people can exist in various periods of time (e.g. , recognizing oneself at different ages, autobiographic stories, etc.). Likewise, they learn that some events may be used as temporal landmarks to explain what happened before or after. Linguistically speaking, this all enables the use of the verbal tenses system and of various temporal adverbials. Finally, from 5/6 to 10/11 years, children slowly perceive time as a human creation. For instance, they grasp what durations or speeds are, or that time can be expressed using historical events. Hence, they start being able to use units (minutes, hours, months, centuries, etc.), and to compare situations along extended periods of time. The overall conclusion is that the younger the children, the more simple verbal tenses and the less diverse and complex temporal connectors and adverbials (Vion and Colas, 1999).

Emotions are also reported as a situational dimension that contributes to establishing and maintaining the coherence of facts in a text (Mouw et al., 2019). In particular, three types of emotional information linked to developmental stages in comprehension are distinguished (Blanc, 2010): lexicalized (e.g. "Mary is afraid"), behavioral (e.g. "Mary bursts into tears") and suggested (e.g. "The wolf arrives") emotions. In particular, emotional lexicon has long been of interest to psycholinguists, for instance in French with the lexicons GALC (Scherer, 2005) and EMOTAIX (Piolat and Bannour, 2009). Studies have also shown the importance of distinguishing, in language addressed to children, the basic emotions (joy, anger, sadness, fear) from more complex ones (guilt, pride, etc.), which are acquired around the age of 10 (Davidson, 2006).

2.2. Learning to Read

As soon as children start learning to read, the question of how well they read is essential to model their ability to understand what they read. Among the different models of the learning to read process, the Frith's model seems to be one of the most widely admitted (Frith, 1985). This model argues that reading is acquired through 3 main stages, namely the logographic, alphabetical, and orthographic stages. The logographic stage, between the ages of 5 and 6, refers to the faculty of recognizing the drawing of a word rather than deciphering it. This first stage is very limited because it allows the memorization of no more than a hundred words. The second one, the alphabetic stage, refers to the faculty of breaking down a word into simpler graphic units (named graphemes), and of converting these graphic units into phonological ones (named phonemes). This stage then gives the possibility to systematically decrypt known or unknown words, with one major difficulty being that the association between graphemes and phonemes is not one-to-one. At the third stage, the orthographic one, a linguistic ability to break down a word into meaningful units (morphemes) can be described.

Limits of Frith's model are that the most recent models tend to refute a discrete evolution of learning to read, preferring an interactive evolution between language learning and reading. It may also be noted that other elements outside the text may affect a child's understanding of the text. For example, the intonation used when reading a text can influence the perception of the text. This phenomenon evolves with age (Aguert et al., 2009).

Considering possible transpositions of this model, the graphical and phonological complexities of words can be associated to the corresponding ages of the Frith's stages.

Finally, it is worth noting that some work propose computer-assisted software to learn to read by stimulating specific aspects taking part in a text's comprehension (Potocki et al., 2013; Beucher-Marsal et al., 2015). This help may focus on the decoding of words (for instance, by splitting it into syllables), by associating words with pictures or with their root phrase (anaphoric links). Beyond the linguistic aspects, the methodological part of these works will be useful if we conduct evaluation campaigns with actual children.

2.3. Measuring the Legibility of a Text

Among related topics, the notion of legibility of a text has been considerably studied over time. Historical formulas were made for English to determine a level of study associated with a text. They are based on lexical and syntactic complexities. For example, the Flesch-Kincaid index (Flesch, 1948) computes frequency ratios on syllables, words and sentences. This index is then mapped to a level of schooling. Many variants exist, usually introducing specific aspects. For example, the Dale-Chall readability formula considers "difficult" words (Dale and Chall, 1948).
In view of the previous remarks on the cognitive and linguistic development of children, it nevertheless seems clear that the only criteria of lexical complexities and syntax are insufficient to state whether a child is able to understand a given text or not. Over the last 20 years, with advances in computer science, including automatic language processing, new approaches have made it possible to consider more criteria and more elaborate method to combine them.

2.4. On the Side of Computational Approaches

There are no works on automatic prediction of age recommendation. However, a number of existing computational studies with methodologies or criteria provide us with some interesting perspectives. They concern either texts simplification for children (De Belder and Moens, 2010; Gala et al., 2018), or acquisition of French as a foreign language (François and Fairon, 2012). (De Belder and Moens, 2010) mainly focuses on lexical simplification (by replacing words with synonyms shorter in size) and syntactic simplification (by decreasing the number of words in a sentence thanks, for instance by removing subordinate links). (François and Fairon, 2012) mentions 46 linguistic criteria (on lexical, syntactic and semantic levels) from which several models are defined to predict levels of readability. The authors then compare the models with each other, as well as with a random classifier. As a result, the most complex methods are the most efficient, even if the rate of good classification tends to cap below 80 per cent. Among the different linguistic criteria, the lexical aspect (more or less complex words) seems to impact the most the prediction's efficiency. Then come the syntactic aspects, already exploited by previous works on readability, and finally the criteria specific to the study of French as a foreign language. Whereas our work shares several aspects with (François and Fairon, 2012), it is different in various others. First, we aim at predicting an actual age, not at classifying into language proficiency grades (of Europe, 2001). This is important since these grades solely focus on linguistic aspects, whereas our purpose also includes developmental and cognitive ones. Secondly, we propose advanced machine learning model (deep neural networks) and introduce or deepen some linguistic features types (embeddings, phonetics, sentiments, syntactic dependencies, etc.—see Section 3.). Finally, our work includes a comparison with predictions made by experts.

3. Features

The state of the art leads us to consider a list of 39 linguistic aspects that may be clues for age recommendation. As detailed below, these aspects are gathered in 10 categories, leading to a global feature vector of 606 real values for each input text.

Embeddings (1 feature of dimension 500)

- Average of the word embeddings. The embeddings, taken from (Fauconnier, 2015), are 500-dimensional features trained for French using skip-grams on the FrWaC corpus (Baroni et al., 2009). As in many NLP tasks, it is meant that the semantic and morphosyntactic information conveyed by these embeddings should help the models.

Lexical information (5 features)

- Mean and standard deviation of log probability of the words in French. Log probabilities have been derived from the language model for French for the speech recognition, trained on types of various types.
- Diversity of lemmas.
- Mean and standard deviation over the frequencies of the words.

Graphy/typography (6 features)

- Mean and standard deviation over the graphical confusability score of the words. To do so, we consider a graphical confusion score $c(x, y)$ between two graphemes x and y. Then, given a word $w = [w_1 \dots w_N]$, the confusability score if computed as the cumulative confusion between each pair of consecutive graphemes in the word, that is: $\sum_{i=1}^{N-1} c(w_i, w_{i+1})$. In practice, the confusion score c is taken from (Geyer, 1977).
- Mean and standard deviation over of the length of the words.
- Ratio of characters (including punctuation marks) against the number of words.
- Ratio of punctuation marks against the number of words.

Morphosyntax (7 features)

- Proportion of the following grammatical classes: verbs, state verbs, names, adjectives, clitics, temporal adverbs. Part of speech tags are generated using Bonsai (Candito et al., 2010).
- Proportion of stop words in a list of 114 words from (Ranks NL, 2019).

Verbal tenses (24 features)

- Number of different verbal tenses
- Proportions of 7 so-called simple tenses: present, past simple, future, imperfect, subjunctive present, conditional present, infinitive.
- Proportions of 7 composed tenses: compound past (*passé composé*), past past (*passé antérieur*), future past (*futur antérieur*), more than perfect (*plus que parfait*), subjunctive past, past conditional, past infinitive.
- Number of different temporal systems: past, present, future.

- Proportions of conjugated verbs for each of the 3 temporal systems: past, present, future.
- Proportion of compound tenses.
- Proportion of simple tenses.
- Proportion of each mode: infinitive, indicative, subjunctive.

Genders and numbers (5 features)

- Proportion of conjugated verbs in the first person.
- Proportion of conjugated verbs in the second person.
- Proportion of conjugated verbs in the third person.
- Proportion of conjugated verbs in the singular form.
- Proportion of conjugated verbs in the plural form.

Syntactic dependencies (8 features)

- Number of words per sentence
- Average distances (word count) between a word and its dependencies. Dependency parsing is achieved using Bonsai (Candito et al., 2010).
- Maximum distances (word count) between a word and its dependencies.
- Mean and standard deviation of dependencies per word (words that points to a given word).
- Mean and standard deviation of the distances between each word and the words to which it points.
- Depth of the dependency tree.

Logical connectors (16 features)

- Proportion of logical connectors for each of the following types: addition; time; goal; cause; comparison; concession; conclusion; condition; consequence; enumeration; explanation; illustration; justification; opposition; restriction; exclusion. Since the way to gather connectors in categories, varies across papers, the categorization used is a consensus of all of them.

Phonetics (9 features)

- Number of phonemes in the sentence, as generated using the grapheme-to-phoneme convertor of eSpeak (Duddington, Jonathan, 2014).
- Number of different phonemes in the sentence.
- Frequency of the phonemes over the whole sentence.
- Mean and variance of the phonetic ordinariness scores of the words. The ordinariness score is computed as the average probability of appearance of each phoneme in French, as given in (Gromer and Weiss, 1990).
- Mean and variance of the word-based diversity of the phonemes.
- Mean and variance of the number of phonemes per word.

Sentiments/emotions (26 features)

- Score of subjectivity as used in the sentiment classifier TextBlob (Loria, 2018)
- Score of polarity (still using TextBlob).
- Proportion of words identified as trigger for a predefined set of 24 emotions: neutral, admiration, love, appeasement, daring, anger, behavior, guilt, disgust, displeasure, desire, embarrassment, empathy, pride, impassibility, inhumanity, jealousy, joy, contempt, unspecified, pride, fear, resentment, surprise, sadness. This list is a refinement of the EMOTAIX dictionary (Piolat and Bannour, 2009).

4. Models

The objective of the current work is to predict from which age an input text can be understood. Whereas this prediction can be barely seen as a regression problem, the definition of ages is more difficult. First, it is important to highlight that our focus is on people with no learning difficulties, e.g. dyslexia, attention problems or mental retardation. This being said, the recommended age can be modeled in different ways. Most obviously, it can be seen as a real value. However, it is known that all children do not develop their skills at the exact same age. It may thus be a better idea to predict a range of ages from which a text can be properly understood. In a third approach, it may be interesting to firstly differentiate texts for children from those for "adults", i.e., people who master all aspects of a text and thus can presumably understand any text[1]. Then, texts for children could be studied more deeply and associated this ages or age ranges.

In any case, the definition of the age from which readers can be viewed as adult is difficult. In the literature, it is admitted that reading comprehension still improves in young adults (18-30 years old). For instance, improvements are related to the abilities to connect with individual knowledge and to recognize types of problems addressed in a text and anticipate how they will be addressed (Baker, 1989). However, since the current work can be considered as a milestone towards more sophisticated analyses of age recommendation, we decided to set the boundary between child and adults readers to 16 years old or, if considering a range of ages, the range of 14-18. This age range coincides with the period when people are supposed to be in high school.

Hence, in this paper, we propose 3 types of models, all implemented as feed-forward fully connected neural networks and illustrated in Figure 1. All of them take a 606-dimensional vector of global features as input. Their details are as follows.

(A) The first model is a standard regression model. The first layer is of dimension 606. The other hyperparameters (number and size of the other hidden layers, activation function, dropout) will be detailed in Section 6.1., and copied for the other models.

[1] With the exception of advanced knowledge which may require expertise in a specific domain (e.g. in mechanics, information, etc.)

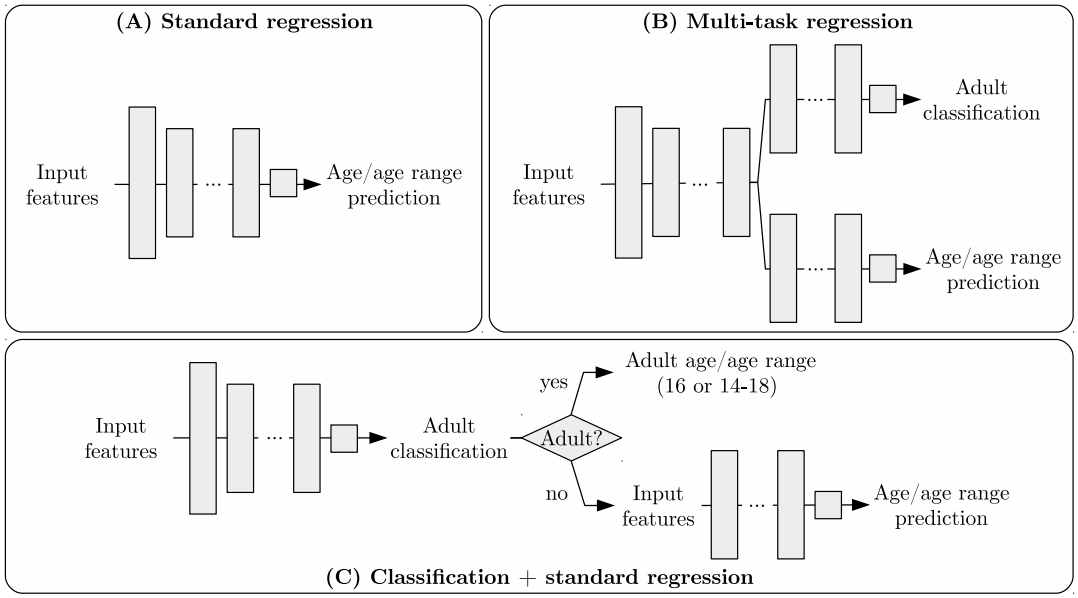

Figure 1: Overview of the 3 types of models. For each of them, the prediction can be either an age or a range of ages from which a text can be read.

(B) The second model is a multi-task model where age prediction is augmented with a binary classification task where inputs shall be judged as dedicated to adults or not ("children"). In addition to hyper-parameters of model A, the number of task-specific layers is investigated in Section 6.1..

(C) The last model is the sequence of a classifier and, if the class is predicted as "children", a regression model. The idea is that there is no need to predict values is the input is considered as "adult" since the age and age ranges are known, i.e., 16, and 14-18, respectively. The hyper-parameters used for model C are the same as those of model A. The regression model of model C is trained on a dataset restricted to children texts only.

For each model, we consider two variants: one which predicts a single minimal age; a second which first predicts an age range, that is two bounds for the minimal age, and then computes the average of this two bounds. The model are evaluated in terms of absolute difference between the predicted age and the expected one (ground truth). Hence, on a corpus, the performance will be given in term of Mean Average Error (MAE).

5. Data

To our knowledge, there is no publicly available dataset for our problem, i.e., a set of texts associated with a target age or age range. Furthermore, most of the content dedicated to children is edited by companies, usually magazines. This brings problem of public access and of document structures which are complex to parse (multiple columns, figures, floating texts, etc.). Finally, one has to highlight that there exist encyclopedias dedicated to children, e.g. Wikimini[2] or Vikidia [3]. However, the texts in these resources may be written by children. Hence, we did not use these because we consider that this is an important bias as writing and understanding skills are different.

To experiment the different models, we have collected a set of 632 texts, among which 543 are for children aged between 0 and 14, and the 89 others are considered as being for adults. Texts for children come from tales, novels, magazines, and newspapers. They come along the editors' or authors' indication in the form of an age range A-B, where A and B are the lower and upper bound for age recommendation. A single age recommendation $\frac{A+B}{2}$ is derived from these age ranges. As This single age is considered as the ground truth in the remainder of the paper. Texts for adults are of similar types as children texts. Particular attention has been paid to include text that children would have difficulties to understand, e.g. novels with a formal language, Wikipedia and newspaper articles about advanced topics (capitalism, genetics, diplomatic issues, etc.).

All texts are split into sentences, each being associated with the age range of the parent text. The corpus is made of 30K sentences and about 446K words. They are partitioned into training, development, and test sets according to the distribution 60/20/20 %. To measure the intra-text dependency of the models, a part of the test is composed of sentences coming from texts that are not seen at all in the training and development set. The objective is to test the model on texts which are indisputably different from what the model is used to see. Furthermore, the distributed of age ranges in this "unseen" set is different (more uniform), even including new age ranges. All the CSV files can be found on `http://texttokids.parisnanterre.fr`. Due to copyright issues, all the texts could not be released. New versions of the corpus will be published in the future, including texts when possible.

On the training set, the average age is 10.26, while the average age range is 8.33-12.19. This is about similar for the

[2] `fr.wikimini.org`

[3] `fr.vikidia.org`

Sigmoid

Size of each layer \ Number of hidden layers	1	2	3	4	5	6	7	8	9	10
10	2.22	2.17	10.22	10.22	2.31	10.22	2.28	10.22	2.42	10.22
20	2.20	2.16	2.24	2.25	10.22	2.22	10.22	10.22	2.32	2.43
50	2.25	2.26	2.22	2.27	2.21	2.22	10.22	10.22	10.22	2.24
100	10.22	2.19	2.18	2.26	10.22	2.16	10.22	2.20	2.23	2.43
200	2.34	2.20	10.22	2.23	**2.07**	10.22	10.22	2.21	2.21	10.22
400	2.36	10.22	2.15	10.22	2.13	10.22	2.26	10.22	3.40	10.22

tanh

Size of each layer \ Number of hidden layers	1	2	3	4	5	6	7	8	9	10
10	2.48	2.35	2.33	2.28	2.35	2.34	2.37	2.36	2.33	2.39
20	2.29	2.19	2.28	2.35	2.30	2.37	2.30	2.35	2.28	2.32
50	2.26	2.19	2.16	2.23	2.20	2.27	2.24	2.29	2.25	2.26
100	2.26	2.20	2.16	2.19	2.27	2.29	2.26	2.22	2.21	2.25
200	2.26	**2.14**	2.21	2.19	2.23	2.42	2.37	2.50	6.51	5.78
400	2.25	2.16	2.20	2.20	2.29	2.32	2.41	6.04	5.87	5.87

ReLU

Size of each layer \ Number of hidden layers	1	2	3	4	5	6	7	8	9	10
10	2.56	2.58	2.63	2.63	2.63	2.66	10.22	2.59	10.22	10.22
20	2.52	2.57	2.52	2.52	2.52	2.44	2.54	10.22	2.44	10.22
50	2.49	2.53	2.39	2.39	2.39	2.10	2.14	2.16	2.25	2.11
100	2.51	2.39	2.27	2.27	2.27	2.08	2.11	2.10	2.10	2.11
200	2.44	2.35	2.24	2.24	2.24	**2.06**	2.11	**2.06**	2.13	2.19
400	2.42	2.31	2.29	2.29	2.29	2.13	2.16	2.15	2.08	2.13

Table 1: Mean average error on the predicted age for various numbers of hidden layers and various layer sizes, for the sigmoid, tanh and ReLU activation functions.

development set. Regarding the test set, they are slightly different (10.05, and 8.20-11.91), especially in the unseen part (9.01, and 7.54-10.48).

As shown in Figure 2, this distribution of the sentences is relatively balanced. A special effort has been made to maximize the size of data while keeping this balance reasonable. In all cases, it was more difficult and time expensive to gather data for early ages since these texts are shorter than those for older children or adults (thus requiring more texts for a same amount of sentences), and they are mainly distributed in a paper form (books). As a consequence, there are very few samples for ages lower than 3.

6. Results

This section presents how models have been trained and tuned (Section 6.1.), before providing final evaluations (Section 6.2.) and discussing the respective influences of the features used (Section 6.3.).

6.1. Tuning (development set)

All models are trained on the training set using the development set to prevent from overfitting. The objective function (loss) is the mean squared error for regression, and the binary cross-entropy for classification. All experiments are run using Adam optimization, with 500 epochs and a batch size of 256 sentences. Programs have been written in Python, using the libraries Keras[4] (Chollet and others, 2018) and TensorFlow[5] (Abadi et al., 2016).

Table 1 reports the MAEs on the development set for various activation functions (sigmoid, tanh, and ReLU), numbers of hidden layers, and sizes for all of these layers. Overall, it appears that sigmoid is unstable regarding the convergence of the optimization process, as cell with a black background indicate. Dropout has been tested but led to worse results. Then, the best results are reported with ReLU and 6 hidden layers of 200 units, without any dropout. This setting is maintained for the rest of the experiments.

Table 2 presents results on the development set for models A, B and C. For model B, the trade-off between shared and task-specific layers is tuned. Results are reported for models either considering direct predictions of the recommended ages or predicting an age range first. In this case, MAEs for each bound of the range are also reported. Accuracy for the adult-children classification task are provided for models B and C. Furthermore, all models are compared with the naive approach which consists in always predicting the mean values observed in the training set (see Section 5.). Overall, all models clearly outperform this naive approach, with MAEs and accuracies spanning between 2.01 and 2.00. The best results are achieved by model B for both age and "age range" strategies, but with two different topologies (either 3 or 4 task-specific layers). Finally, it appear that these two strategies do not seem to lead to significant differences.

		MAE Age	Acc.	MAE Lower bound age	MAE Upper bound age	MAE Avg. of age range	Acc.
Naive		3.44	74.7%	3.59	3.30	3.44	74.7%
A		2.06	–	2.12	2.08	2.09	–
B (# of task-specific hidden layers)	1	2.11	84.8%	2.08	2.03	2.11	85,2%
	2	2.13	84.4%	2.15	2.11	2.13	84,6%
	3	2.14	84.2%	**2.02**	**1.99**	**2.00**	**85,3%**
	4	**2.01**	84.7%	2.12	2.16	2.14	84,6%
	5	2.13	83.7%	2.10	2.03	2.06	84,3%
	6	2.05	85.4%	2.19	2.10	2.15	84,2%
C		2.12	84.0%	2.11	2.09	2.09	84.2%

Table 2: Mean average errors and accuracy on the development set for the naive, regression (A), multi-task (B), and classification+regression (C) approaches. Results are given when considering various numbers of various trade-offs between shared and and task-specific hidden layers.

6.2. Evaluation (test set)

To confirm results on the development set, results are given for the best settings of models A, B, and C in Table 3. In complement, a new variants of model B is tested. Similarly to model C, the idea is to force the prediction to the adult values (16, or 14-18) as soon as the text is classified as adult. This variant is referred to as "with forcing". Again, all models are from far better than the naive approach and lead to close results. While model B was the best on the dev set, it seems that the difference was not significant. Further-

[4] keras.io

[5] www.tensorflow.org

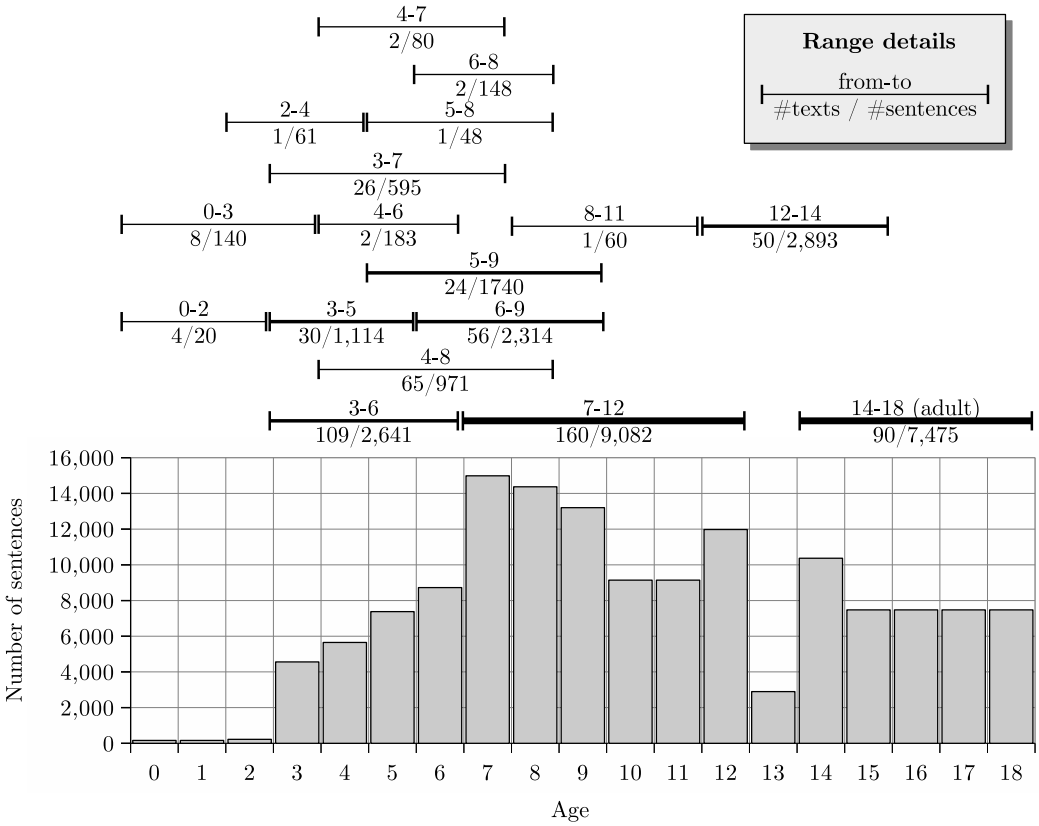

Figure 2: Distribution of data: age ranges present in the corpus (top) and total number of sentences for each age (bottom).

		Mean absolute error on			
		Age	Lower bound age	Upper bound age	Avg. of age range
Naive		3.67	3.74	3.61	3.67
A		**2.24**	**2.26**	2.29	**2.27**
B	w/o forcing	2.26	2.27	2.30	**2.27**
	w/ forcing	2.27	**2.26**	2.30	**2.27**
C		2.31	2.29	2.33	2.30

Table 3: Mean average errors on the test set for the naive, regression (A), multi-task (B), and classification+regression (C) approaches. For models B and C, results are given with or without forcing ages when the adult class is predicted.

more, the forcing trick does not improve the results.

To analyze the quality of the automatic age recommendations, series of experiments were conducted with psycholinguists. Three experts of children's development were given the unseen part of the test and had to propose an age range, centered on their age recommendation, between 0 and 18 years. Annotations have been performed either on the whole texts from the unseen part, or on 80 random sentences, i.e., taken without any context. Table 4 shows this results compared against the naive approach and our best model (model A). The results of the experts are given individually or averaged for each sentence or text. On sentences (top table), it appears that the experts also perform better that the naive approach, and that our model predicts

	Mean absolute error on			
	Age	Lower bound age	Upper bound age	Avg. of age range
80 sentences				
Naive	4.46	4.29	4.63	4.46
Best model	**2.70**	**2.53**	**2.65**	**2.57**
Expert 1	3.14	2.95	3.45	3.14
Expert 2	3.38	3.48	3.39	3.38
Expert 3	3.07	2.93	3.54	3.07
Mean of experts	2.88	2.86	3.05	2.88
20 texts				
Naive	4.57	4.32	4.83	4.57
Best model (per sentence)	3.13	3.07	3.35	3.18
Best model (per text)	**2.39**	**2.51**	**2.57**	**2.53**
Expert 1	2.60	2.60	2.80	2.60
Expert 2	3.50	3.80	3.30	3.50
Expert 3	2,70	2,90	2,60	2,70
Mean of experts	2.95	3.19	2.81	2.95

Table 4: Mean average errors on the unseen part of the test set for the naive approach, our best model, and the experts annotations.

recommendations which are closer to the ground truth than those of the experts.

On the contrary, as soon as more context is given with the full texts (bottom table), the expert provide better recommendations than the sentence-level predictions (line "per

Léa, l'oiseau et Lapin regardent Renard. (Lea, the bird and Rabbit look at Fox.)
Ground truth: 3 (2-4)
Best model: 2.9
Mean of experts: 4.0 (3.7-4.3)

Il jouait de ses instruments jour et nuit. (He played his instruments day and night.)
Ground truth: 9.5 (8-11)
Best model: 7.0
Mean of experts: 6.8 (6.0-7.7)

Pourtant, certains ont développé une technique d'écriture qui permet de la contourner. (However, some have developed a writing technique that allows to circumvent it.)
Ground truth: 13 (12-14)
Best model: 9.5
Mean of experts: 9.8 (8.7-11.0)

Ce qu'on entendait au faubourg Saint Jacques, pendant la nuit du mardi gras au mercredi des cendres, dans la cour d'un pharmacien droguiste. (What was heard in the Faubourg Saint Jacques during the night of Mardi Gras and Ash Wednesday in the courtyard of a druggist pharmacist.)
Ground truth: 16 (14-18)
Best model: 16.0
Mean of experts: 12.2 (11.0-13.3)

Table 5: Sample sentences with their ground truth age and age range, along with automatic and human predictions.

sentence"). However, in this configuration, we studied the possibility to aggregate the recommendations returned for each constituent sentence. To do so, predictions are averaged. The resulting performance of our model, given by the line "per text", significantly improves, finally bringing to better predictions than those of the experts.

A detailed analysis of the results highlights that the difference between automatic and human prediction is large. On sentences, the absolute difference is 2.67, while it is 1.15 on full texts. This is surprising with respect to the differences previously observed on MAEs (respectively, 0.18 and 0.56). In our opinion, after inspecting the distribution of the predictions, the main reason is that the experts use lower interval of values than the model. Most recommendation made by the experts are between 4 and 12, whereas the model maps its values into a larger domain, between 1.5 and 16 as observed at the training time. This phenomenon is illustrated by the examples of Table 5.

6.3. Impact of the Feature Groups

Finally, we have experimented the training of our best model by discarding a group of features (see Section 3.) or using only one group. The results, as given in Table 6, show that the embeddings are clearly preponderant in the decision. Then, apart from the embeddings, no group can predict better values than the naive approach when used alone. While this once again shows that embeddings are a solution to many NLP problems, it does not provide any feedback to the psycholinguistic side of age recommendation. Further studies on features will thus be conducted to provide a better view of their respective influences.

Feat. group (X)	All\X	Δ	Only X	Δ
Naive	3.67	–	–	–
All	2.24	–	–	–
Embeddings	3.26	+1.02	2.32	+0.09
Lexical	2.25	+0.01	3.70	+1.46
Graph./typograph.	2.25	+0.01	3.79	+1.55
Morphosyntax	2.26	+0.02	3.86	+1.62
Verbal tenses	2.29	+0.06	3.66	+1.43
Genders, numbers	2.25	+0.02	3.62	+1.39
Syntactic dep.	2.29	+0.05	3.94	+1.70
Logical connectors	2.31	+0.08	3.64	+1.41
Phonetics	2.39	+0.16	3.85	+1.61
Sentiments	2.32	+0.08	3.67	+1.43

Table 6: Impact of each group of features (X) on mean absolute errors by considering each one alone (All\X), or excluding it (Only X). Difference with the best model are given (Δ).

7. Conclusion and Perspectives

In this paper, we have investigated the problem of age recommendation for texts, which is an original task in NLP. Several neural models and strategies have been proposed, and results have been compared with the performance of psycholinguists. These results show that our model's predictions are better than those of experts. Furthermore, while relying on a strong assumption that all sentences of a text can be seen as all dedicated to a unique age range, our results on the aggregation of sentence-based results are clearly encouraging. This demonstrates the viability of the approach and calls for further investigations.

However, we are aware that these results should be taken with caution as various aspects bring uncertainty in the experimental process, leading to methodological perspectives. In particular, data annotations are provided from editors and authors. One may wonder how accurate these annotations are since they can sometimes integrate editorial guidelines, e.g. specific age range for some book collections. Then, considering mean absolute errors with one target age is probably too hard. It could be interesting to integrate the age range in the evaluation. For instance, one may consider the error as null as soon the prediction fits inside the reference age range. Finally, it would be interesting to correlate the results with an *in situ* evaluation campaign with children. This is planned in the next few months as part of the project in which the current work has been conducted. Regarding technical aspects, the main perspective is to try using massive corpora (potentially) written by children (e.g. Wikimini and Vikidia) as bootstraps for more advanced neural networks, especially recurrent ones.

8. Acknowledgements

This work has been partially funded by the French National Research Agency (ANR) through the projects TREMoLo and TextToKids. Furthermore, we would like to thank Nathalie Blanc, Aliyah Morgenstern and Christophe Parisse for their contribution as psycholinguists.

9. Bibliographical References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.

Aguert, M., Bernicot, J., and Laval, V. (2009). Prosodie et compréhension des énoncés chez les enfants de 5 à 9 ans. *Enfance*, 2009, 09.

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review*, 1(1).

Beucher-Marsal, C., Charles, F., and Le Hénaff, C. (2015). Improving literacy skills and differentiating learning speed among primary school children through a computer-assisted learning tool. *The International Journal of Literacies*.

Blanc, N. (2010). La compréhension des contes entre 5 et 7 ans: Quelle représentation des informations émotionnelles? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 64(4).

Candito, M., Nivre, J., Denis, P., and Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for french. In *Proceedings of the International Conference on Computational Linguistics*. `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`.

Chollet, F. et al. (2018). Keras: The python deep learning library. *Astrophysics Source Code Library*.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*.

Davidson, D. (2006). The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, 30(3).

De Belder, J. and Moens, M.-F. (2010). Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, 01.

Duddington, Jonathan. (2014). espeak.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3).

François, T. (2015). When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2).

François, T. and Fairon, C. (2012). An "AI readability" formula for french as a foreign language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Frith, U. (1985). Beneath the surface of developmental dyslexia. *In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.),Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*.

Gala, N., Francois, T., Javourey-Drevet, L., and Ziegler, J. C. (2018). Text simplification, a tool for learning to read. *Langue française*, (199).

Gathercole, S. (1999). Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, 3, 12.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, 22(5), Sep.

Gromer, D. and Weiss, M. (1990). *Lire, tome 1 : apprendre à lire*. Armand Colin.

Hickmann, M. (2012). Diversité des langues et acquisition du langage: espace et temporalité chez l'enfant. *Langages*, (4).

Loria, S. (2018). Textblob.

Mouw, J. M., Van Leijenhorst, L., Saab, N., Danel, M. S., and van den Broek, P. (2019). Contributions of emotion understanding to narrative comprehension in children and adults. *European Journal of Developmental Psychology*, 16(1).

of Europe, C. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Piolat, A. and Bannour, R. (2009). An example of text analysis software (emotaix-tropes) use: The influence of anxiety on expressive writing. *Current psychology letters. Behaviour, brain & cognition*, 25(2, 2009).

Potocki, A., Ecalle, J., and Magnan, A. (2013). Effects of computer-assisted comprehension training in less skilled comprehenders in second grade: A one-year follow-up study. *Computers & Education*, 63.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, 44(4).

Tartas, V. (2010). Le développement de notions temporelles par l'enfant. *Développements*, 4.

Vion, M. and Colas, A. (1999). L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements). In *Proceedings of the Conference of the International Association for the Study of Child Language*.

10. Language Resource References

Baroni, Marco and Bernardini, Silvia and Ferraresi, Adriano and Zanchetta, Eros. (2009). *The WaCky wide web: a collection of very large linguistically processed web-crawled corpora*. Springer.

Fauconnier, Jean-Philippe. (2015). *French Word Embeddings*. `http://fauconnier.github.io`.

Ranks NL. (2019). *List of French stop words*. `https://www.ranks.nl/stopwords/french`.